2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

SVD truncation schemes for fixed-size kernel models

Ricardo Castro, Siamak Mehrkanoon, Anna Marconato, Johan Schoukens and Johan A. K. Suykens

***Abstract*—In this paper, two schemes for reducing the effective number of parameters are presented. To do this, different versions of Fixed-Size Kernel models based on Fixed-Size Least Squares Support Vector Machines (FS-LSSVM) are employed. The schemes include Fixed-Size Ordinary Least Squares (FS-OLS) and Fixed-Size Ridge Regression (FS-RR) with their respective truncations through Singular Value Decomposition (SVD). When these schemes are applied to the Silverbox and Wiener-Hammerstein data sets in system identification, it was found that a great deal of the complexity of the model could be reduced in a trade-off with the generalization performance.**

I. Introduction

When evaluating modeling techniques several performance criteria can be used. Normally, performance based on an error cost function is evaluated on a test set as this illustrates the generalization performance of the model. However, there might be other desirable characteristics of the models. For instance, where control is the goal of the identified model, a low complexity is also desirable by itself besides a good generalization capacity [12].

The authors acknowledge support from Research Council KUL: GOA/10/09 MaNet , PFV/10/002 (OPTEC), several PhD/postdoc and fellow grants; Flemish Government: IOF: IOF/KP/SCORES4CHEM FWO: PhD/postdoc grants, projects: G.0377.12 (Structured systems), G.083014N (Block term decompositions), G.088114N (Tensor based data similarity) IWT: PhD Grants, projects: SBO POM, EUROSTARS SMART iMinds 2013 Belgian Federal Science Policy Office: IUAP P7/19 (DYSCO, Dynamical systems, control and optimization, 2012-2017) EU: FP7-SADCO (MC ITN-264735), ERC ST HIGHWIND (259 166), ERC AdG A-DATADRIVE-B (290923) COST: Action ICO806: IntelliCIS)

This work was also supported in part by the Fund for Scientific Research (FWO-Vlaanderen), by the Flemish Government (Methusalem), by the Belgian Government through the Inter university Poles of Attraction (IAP VII) Program and the ERC Advanced Grant SNLSID.

Ricardo Castro and Siamak Mehrkanoon are with the Department of Electrical Engineering - ESAT, STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven, B-3001 Leuven, Belgium (e-mail: ricardo.castro, siamak.mehrkanoon @esat.kuleuven.be).

Johan A. K. Suykens is with the Department of Electrical Engineering-ESAT, STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, and iMinds Future Health Department, KU Leuven, B-3001 Leuven, Belgium (e-mail: johan.suykens@esat.kuleuven.be).

Anna Marconato and Johan Schoukens are with Dept. ELEC, Vrije Universiteit Brussel, Brussels, Belgium. Email: anna.marconato@vub.ac.be

For assessing the generalization performance of trained models without the use of validation data, various criteria have been developed. Such criteria take the general form of a prediction error (PE) which consists of the sum of two terms, namely PE = training error + complexity term. The complexity term represents a penalty growing with the number of free parameters in the model. Clearly, when the model is too simple it will be penalized by the residual error, but if it is too complex, it will be penalized by the complexity term. The minimum value for the criterion is given by a trade-off between the two terms [1].

In [14] Moody generalized such criteria to deal with non-linear models and to allow for the presence of a regularization term through the generalized prediction error which includes the effective number of parameters. Other approaches, like the one presented by Vapnik and Chervonenkis in [20] proposed an upper bound on the generalization error with a complexity term depending on the Vapnik-Chervonenkis dimension. Several other different theories with different notions of model complexity have been proposed in literature.

It is well-known that when applying regularization then instead of the number of parameters, the effective number of parameters is a more suitable notion then for model complexity. Also within support vector machines and kernel-based models the use of regularization is common [19]. Within the context of this paper we will consider different versions of fixed-size kernel models related to fixed-size least squares support vector machines [19]. We will consider the *effective degrees of freedom* here as the notion for model complexity. The effective degrees of freedom is characterized by the trace of the hat matrix. The studied fixed-size kernel models relate to applying ordinary least squares and ridge regression in the primal, after obtaining a Nyström approximated feature map based on a selected subset of the given data. The resulting kernel models are sparse and the terminology of support vectors is used here for the Rényi based selected subset of prototype vectors. The size of the subset controls the degree of sparsity of the fixed-size kernel model.

Through this work, SVD truncation schemes for the fixed-size kernel models are investigated. It will be illustrated that even though these truncation schemes are

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

not suited to further improve the generalization performance, the effective degrees of freedom can be greatly reduced. This realizes a reduction of the complexity of the resulting models and in this way, the resulting model can keep a fairly good generalization performance while at the same time getting a reduced complexity.

In this work scalars are represented in lower case, bold lower case is used for vectors and capital bold stands for matrices. e.g. x is a scalar, $\boldsymbol{x}$ is a vector and $\boldsymbol{X}$ is a matrix.

The work is organized as follows: In section II function estimation using LS-SVM and Fixed-Size LS-SVM is explained. In section III, the SVD truncation schemes employed are presented and the concept of effective degrees of freedom is explained. Also, some practical considerations for the implementation done are exposed. In section IV the Silverbox and Wiener-Hammerstein data sets are presented and the results found for the application of SVD truncation schemes are illustrated. These results are discussed on section V. Finally, in section VI the conclusions are given.

II. Fixed-Size LS-SVM

In this section, the different methods used in this work will be exposed. First, a brief introduction to function estimation through LS-SVM is presented. Then, the concept of *effective degrees of freedom* is explained. Finally, the considerations to make an estimation in the primal space are given.

A. Function estimation using LS-SVM

The framework of LS-SVM is given by a primal-dual formulation. Given the data set $\{\boldsymbol{x}_i, y_i\}_{i=1}^N$, the objective is to find a model

$$\hat{y} = \boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{x}) + b \tag{1}$$

where $\boldsymbol{x} \in \mathbb{R}^n$, $\hat{y} \in \mathbb{R}$ denotes the estimated value, and $\boldsymbol{\varphi}(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}^{n_h}$ is the feature map to a high dimensional (possibly infinite) space.

An optimization problem is then formulated [19]:

$$\begin{array}{ll} \min\limits_{\boldsymbol{w},b,e} & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \frac{\gamma}{2}\sum\limits_{i=1}^{N} e_i^2 \\ \text{subject to} & y_i = \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_i) + b + e_i,\ i = 1, ..., N. \end{array} \tag{2}$$

Through the use of Mercer's theorem [13], the entries of the kernel matrix $\boldsymbol{\Omega}_{i,j}$ can be represented by $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{\varphi}(\boldsymbol{x}_i)^T\boldsymbol{\varphi}(\boldsymbol{x}_j)$ with $i, j = 1, ..., N$. Note then that $\boldsymbol{\varphi}$ does not have to be explicitly known as this is done implicitly through the positive definite kernel function. In this case, the radial basis function kernel (RBF kernel) was used i.e. $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_2^2 / \sigma^2)$ where σ is a tuning parameter.

From the Lagrangian $\mathcal{L}(\boldsymbol{w}, b, e; \alpha) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\sum_{i=1}^N e_i^2 - \sum_{i=1}^N \alpha_i(\boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_i) + b + e_i - y_i)$ with $\alpha_i \in \mathbb{R}$ the Lagrange multipliers, the optimality conditions for this formulation are:

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = 0 \rightarrow \boldsymbol{w} = \sum_{i=1}^N \alpha_i \boldsymbol{\varphi}(\boldsymbol{x}_i) \\ \frac{\partial \mathcal{L}}{\partial b} = 0 \rightarrow \sum_{i=1}^N \alpha_i = 0 \\ \frac{\partial \mathcal{L}}{\partial e_i} = 0 \rightarrow \alpha_i = \gamma e_i, i = 1, ..., N \\ \frac{\partial \mathcal{L}}{\partial \alpha_i} = 0 \rightarrow y_i = \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_i) + b + e_i, i = 1, ..., N. \end{cases} \tag{3}$$

By elimination of $\boldsymbol{w}$ and e_i the following linear system is obtained:

$$\left(\begin{array}{c|c} 0 & \mathbf{1}_N^T \\ \hline \mathbf{1}_N & \boldsymbol{\Omega} + \frac{1}{\gamma}\boldsymbol{I}_N \end{array}\right)\left(\frac{b}{\boldsymbol{\alpha}}\right) = \left(\frac{0}{\boldsymbol{y}}\right) \tag{4}$$

with $\boldsymbol{y} = [y_1, ..., y_N]^T$, $\boldsymbol{\alpha} = [\alpha_1, ..., \alpha_N]^T$. The resulting model is then:

$$\hat{y}(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i) + b. \tag{5}$$

B. Effective degrees of freedom for LS-SVM

The number of model parameters is not a very good indicator of the complexity as it is not a suitable measure for techniques using regularization such as in Support Vector Machines[19]. A possible alternative is the effective degrees of freedom. The effective degrees of freedom can be calculated through the trace of the hat matrix $\boldsymbol{H}$ (also known as the smoother matrix) [5], [11] and [18]. $\boldsymbol{H}$ comes from the expression $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$. For further insight about the effective degrees of freedom see [1], [10], [14] and [18].

For LS-SVM, the hat matrix can be calculated as follows [2]. From (4) and (5) one has:

$$\hat{\boldsymbol{y}} = \boldsymbol{\Omega}\hat{\boldsymbol{\alpha}} + \mathbf{1}_N\hat{b} \tag{6}$$

with

$$\begin{cases} \hat{\boldsymbol{\alpha}} &= \left(\boldsymbol{\Omega} + \frac{\boldsymbol{I}_N}{\gamma}\right)^{-1}\left(\boldsymbol{y} - \mathbf{1}_n\hat{b}\right) \\ \hat{b} &= \frac{\mathbf{1}_N^T\left(\boldsymbol{\Omega} + \frac{\boldsymbol{I}_N}{\gamma}\right)^{-1}\boldsymbol{y}}{\mathbf{1}_N^T\left(\boldsymbol{\Omega} + \frac{\boldsymbol{I}_N}{\gamma}\right)^{-1}\mathbf{1}_N}. \end{cases} \tag{7}$$

Let c and $\boldsymbol{Z}$ be defined as:

$$\begin{array}{rcl} c &=& \mathbf{1}_N^T\left(\boldsymbol{\Omega} + \frac{\boldsymbol{I}_N}{\gamma}\right)^{-1}\mathbf{1}_N \\ \boldsymbol{Z} &=& \boldsymbol{\Omega} + \frac{\boldsymbol{I}_N}{\gamma}. \end{array} \tag{8}$$

Let $\boldsymbol{J}_N$ be defined as a square matrix of size $N \times N$ where all elements are equal to 1. This leads to the hat matrix $\boldsymbol{H}$:

$$\boldsymbol{H} = \left[\boldsymbol{\Omega}\left(\boldsymbol{Z}^{-1} - \boldsymbol{Z}^{-1}\frac{\boldsymbol{J}_N}{c}\boldsymbol{Z}^{-1}\right) + \frac{\boldsymbol{J}_N}{c}\boldsymbol{Z}^{-1}\right]. \tag{9}$$

C. Estimation in primal space using FS-LSSVM

Usually, the feature map should not be explicitly known when solving in the dual. This is the case for the RBF kernel for which the feature map is infinite dimensional [20]. In order to be able to work in the primal space, it is required that either the feature map $\boldsymbol{\varphi}$ is explicitly known and it is finite dimensional (e.g. linear kernel case) or an approximation to $\boldsymbol{\varphi}$ is acquired. This can be achieved through an eigenvalue decomposition of the kernel matrix $\boldsymbol{\Omega}$ with entries $K(\boldsymbol{x}_k, \boldsymbol{x}_l)$. Given the integral equation $\int K(\boldsymbol{x}, \boldsymbol{x}_j)\phi_i(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} = \lambda_i\phi_i(\boldsymbol{x}_j)$ with λ_i and ϕ_i the eigenvalues and eigenfunctions related to the kernel function respectively for a variable $\boldsymbol{x}$ with probability distribution $p(\boldsymbol{x})$. The following expression can be written then [4], [3] and [6]:

$$\hat{\boldsymbol{\varphi}}(\boldsymbol{x}) = \left[\sqrt{\lambda_1}\phi_1(\boldsymbol{x}), \sqrt{\lambda_2}\phi_2(\boldsymbol{x}), ..., \sqrt{\lambda_{n_h}}\phi_{n_h}(\boldsymbol{x})\right]^T. \tag{10}$$

Through the Nyström method [15] and [21], an approximation to the integral equation is obtained by means of the sample average determining an approximation to ϕ_i leading to

$$\frac{1}{N}\sum_{k=1}^{M} K(\boldsymbol{x}_k, \boldsymbol{x}_j)\boldsymbol{u}_{ik} = \lambda_i^{(s)}\boldsymbol{u}_{ij} \tag{11}$$

where $\lambda_i^{(s)}$ and $\boldsymbol{u}_i$ are the sample eigenvalues and eigenvectors respectively.

A finite dimensional approximation $\hat{\boldsymbol{\varphi}}_i(\boldsymbol{x})$ can be computed for any point $\boldsymbol{x}^{(v)}$ through

$$\begin{aligned} \hat{\boldsymbol{\varphi}}_i(\boldsymbol{x}^{(v)}) &= \frac{1}{\sqrt{\lambda_i^{(s)}}}\sum_{k=1}^{M} \boldsymbol{u}_{ki}K(\boldsymbol{x}_k, \boldsymbol{x}^{(v)}) \\ \text{with } i &= 1, \ldots, M. \end{aligned} \tag{12}$$

This approximation can then be used in the primal to estimate $\boldsymbol{w}$ and b.

For large scale problems, a subsample of M datapoints (with $M \ll N$) could be selected to compute $\hat{\boldsymbol{\varphi}}$ together with estimation in the primal. This is known as *Fixed-Size Least Squares Support Vector Machines* (*FS-LSSVM*) [19]. Criteria as entropy maximization has been used to select appropriate M datapoints instead of a merely random approach. In this case, Rényi's entropy H_R is used [8]:

$$H_R = -\log \int p(\boldsymbol{x})^2 d\boldsymbol{x}. \tag{13}$$

The higher the entropy found in the subset of M points used, the better this subset will represent the whole data set.

Once the support vectors are selected through Rényi's entropy, the problem in the primal can be represented as

$$\min_{\boldsymbol{w},b} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \frac{\gamma}{2}\sum_{i=1}^{M}(y_i - \boldsymbol{w}^T\hat{\boldsymbol{\varphi}}(\boldsymbol{x}_i) - b)^2 \tag{14}$$

from where the optimal $\boldsymbol{w}$ and b can be extracted directly. Note that given the selection of $M \ll N$, this is a sparse kernel model.

III. SVD TRUNCATION SCHEMES

Once $\hat{\boldsymbol{\varphi}}$ is calculated, the model in primal form is computed according to the techniques described in this section. The particular studied estimation techniques are introduced in this section as well as the effective degrees of freedom (*EDF*) for Fixed-Size Ordinary Least Squares (FS-OLS) and Fixed-Size Ridge Regression (FS-RR).

A. FS-OLS with truncation

After obtaining the optimal M subsample values through Quadratic Rényi Entropy, the training points are projected into the feature space. This projection depends on the dimensionality given by the number of support vectors selected by the user (i.e. M)

$$\hat{\boldsymbol{\Phi}} = [\hat{\boldsymbol{\varphi}}(\boldsymbol{x}_1), ..., \hat{\boldsymbol{\varphi}}(\boldsymbol{x}_{N_{train}})]^T \tag{15}$$

with $\boldsymbol{X}_{train} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_{N_{train}}]$. From this, matrix $\boldsymbol{Q}$ is defined:

$$\boldsymbol{Q} = \hat{\boldsymbol{\Phi}}^T\hat{\boldsymbol{\Phi}}. \tag{16}$$

The $\boldsymbol{Q}$ matrix can be decomposed through SVD resulting in $\boldsymbol{Q} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T$. Given that $\boldsymbol{Q}$ is a positive semi-definite matrix and $\boldsymbol{Q} = \hat{\boldsymbol{\Phi}}^T\hat{\boldsymbol{\Phi}} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T$ with $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{I}$, $\boldsymbol{V}\boldsymbol{V}^T = \boldsymbol{I}$ and $\boldsymbol{S}$ a diagonal matrix with positive diagonal elements, one has $\boldsymbol{Q} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T = \boldsymbol{U}\boldsymbol{S}\boldsymbol{U}^T = \boldsymbol{V}\boldsymbol{S}\boldsymbol{V}^T$.

After decomposing $\boldsymbol{Q}$, the less relevant singular values from $\boldsymbol{S}$ are discarded successively and the reconstructed $\boldsymbol{Q}$ matrix $\hat{\boldsymbol{Q}} = \boldsymbol{U}\hat{\boldsymbol{S}}\boldsymbol{V}^T$ is used in the validation set to determine the best truncation (i.e. how many singular values are discarded).

The *FS-OLS* model estimate with truncation becomes then:

$$\boldsymbol{w}_{OLS_{trun}} = \left(\boldsymbol{U}\hat{\boldsymbol{S}}\boldsymbol{V}^T\right)^{-1}\hat{\boldsymbol{\Phi}}^T\boldsymbol{y}_{train}. \tag{17}$$

Similarly to equation (15):

$$\hat{\boldsymbol{\Phi}}_{val} = \left[\hat{\boldsymbol{\varphi}}(\boldsymbol{x}_1^{val}), ..., \hat{\boldsymbol{\varphi}}(\boldsymbol{x}_{N_{val}}^{val})\right]^T \tag{18}$$

with $\boldsymbol{X}_{val} = [\boldsymbol{x}_1^{val}, ..., \boldsymbol{x}_{N_{val}}^{val}]$. Therefore:

$$\hat{\boldsymbol{y}}_{val_{OLS,trun}} = \hat{\boldsymbol{\Phi}}_{val}\boldsymbol{w}_{OLS_{trun}}. \tag{19}$$

Once the best truncation is found, the system is applied to the test set:

$$\hat{\boldsymbol{y}}_{test_{OLS,trun}} = \hat{\boldsymbol{\Phi}}_{test}\boldsymbol{w}_{OLS_{trun}}. \tag{20}$$

Here, $\hat{\boldsymbol{\Phi}}_{test}$ is defined as:

$$\hat{\boldsymbol{\Phi}}_{test} = \left[\hat{\boldsymbol{\varphi}}(\boldsymbol{x}_1^{test}), ..., \hat{\boldsymbol{\varphi}}(\boldsymbol{x}_{N_{test}}^{test})\right]^T \tag{21}$$

with $\boldsymbol{X}_{test} = [\boldsymbol{x}_1^{test}, ..., \boldsymbol{x}_{N_{test}}^{test}]$.

B. FS-RR with truncation

For the ridge regression technique, $\hat{\boldsymbol{\Phi}}$, $\hat{\boldsymbol{\Phi}}_{val}$, $\hat{\boldsymbol{\Phi}}_{test}$ and $\boldsymbol{Q}$ are calculated in the same way as described in the *FS-OLS* method. However, the formulation changes as follows:

$$\begin{aligned} \boldsymbol{w}_{RR} &= \left(\hat{\boldsymbol{\Phi}}^T\hat{\boldsymbol{\Phi}} + \lambda\boldsymbol{I}\right)^{-1}\hat{\boldsymbol{\Phi}}^T\boldsymbol{y}_{train} \\ &= \left(\boldsymbol{Q} + \lambda\boldsymbol{I}\right)^{-1}\hat{\boldsymbol{\Phi}}^T\boldsymbol{y}_{train} \end{aligned} \tag{22}$$

where λ is the regularization parameter. Truncation of this solution becomes:

$$\begin{aligned} \boldsymbol{w}_{RR_{trun}} &= (\boldsymbol{U}\hat{\boldsymbol{S}}\boldsymbol{U}^T + \lambda\boldsymbol{U}\boldsymbol{U}^T)^{-1}\hat{\boldsymbol{\Phi}}^T\boldsymbol{y}_{train} \\ &= \boldsymbol{U}\left(\hat{\boldsymbol{S}} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{U}^T\hat{\boldsymbol{\Phi}}^T\boldsymbol{y}_{train}. \end{aligned} \tag{23}$$

Once again, the most appropriate λ value is determined by the validation set (i.e. through *linesearch*) and finally, the resulting model is tested on the test set:

$$\hat{\boldsymbol{y}}_{val_{RR,trun}} = \hat{\boldsymbol{\Phi}}_{val}\boldsymbol{w}_{RR_{trun}} \tag{24}$$

and

$$\hat{\boldsymbol{y}}_{test_{RR,trun}} = \hat{\boldsymbol{\Phi}}_{test}\boldsymbol{w}_{RR_{trun}}. \tag{25}$$

For truncation, the same procedure is used as in FS-OLS, however, besides looking for the best λ value, also the best truncation is looked for. This results in a *gridsearch* approach.

C. Effective degrees of freedom

The hat matrix, from where the effective degrees of freedom can be estimated [2], becomes for OLS:

$$\begin{aligned} \boldsymbol{H}_{OLS} &= \hat{\boldsymbol{\Phi}}(\hat{\boldsymbol{\Phi}}^T\hat{\boldsymbol{\Phi}})^{-1}\hat{\boldsymbol{\Phi}}^T \\ \boldsymbol{H}_{OLS_{trun}} &= \hat{\boldsymbol{\Phi}}(\boldsymbol{U}\hat{\boldsymbol{S}}^{-1}\boldsymbol{V}^T)\hat{\boldsymbol{\Phi}}^T. \end{aligned} \tag{26}$$

Similarly, for Ridge Regression and its truncated version:

$$\begin{aligned} \boldsymbol{H}_{RR} &= \hat{\boldsymbol{\Phi}}(\hat{\boldsymbol{\Phi}}^T\hat{\boldsymbol{\Phi}} + \lambda\boldsymbol{I})^{-1}\hat{\boldsymbol{\Phi}}^T \\ \boldsymbol{H}_{RR_{trun}} &= \hat{\boldsymbol{\Phi}}\boldsymbol{U}(\hat{\boldsymbol{S}} + \lambda\boldsymbol{I})^{-1}\boldsymbol{U}^T\hat{\boldsymbol{\Phi}}^T. \end{aligned} \tag{27}$$

IV. Experimental results

In the *FS-LSSVM* it is necessary to specify a subset of M input points to represent the data set reasonably well. For this purpose, the quadratic Rényi entropy is used and an approximation to the feature map is calculated as explained in section II-C. A *gridsearch* approach is used then to tune the values of the tuning parameters λ and σ. The parameters are selected in accordance with the results obtained from evaluating the resulting model on the validation set. The chosen model is finally used on the test set.

Note that the structure of the model will be that of a nonlinear autoregressive model with exogenous input (NARX), where the model relates the current value of a time series with past values of the same series and current and past values of the driving (exogenous) series. A NARX model can be expressed as follows:

$$\hat{y}_t = f(y_{t-1}, y_{t-2}, \ldots, y_{t-p}, u_t, u_{t-1}, u_{t-2}, \ldots, u_{t-p}) \tag{28}$$

where $f(\cdot)$ is some nonlinear function and $\hat{y}_t$ is the estimated value of y. Here y is the variable of interest, u is the external input and p is the number of lags used determining how many past u and y values are included to calculate $\hat{y}$.

In this section, the results obtained by applying the techniques explained in sections II and III under the one-step ahead framework are presented. Also, a description of the data sets used is offered.

A. Silverbox data set

The Silverbox data set was introduced by J. Schoukens, J.G. Nemeth, P. Crama, Y. Rolain and R. Pintelon in [16]. This data set represents an electrical circuit simulating a mass-spring damper system. It is a nonlinear dynamic system with feedback exposing a dominant linear behavior [4].

In Figure 1, the inputs and outputs of the system are depicted. The data set consists of 131072 data points and was split evenly between test, validation and training sets.

B. Wiener-Hammerstein data set

The concatenation of two linear systems with a static nonlinearity in between constitutes an important special class of nonlinear systems known as a Wiener-Hammerstein system [7].

The Wiener-Hammerstein data set was introduced by J. Schoukens, J. Suykens and L. Ljung in [17]. The system modelled is an electronic nonlinear system with a Wiener-Hammerstein structure as shown in Figure 2. There, G_1 is a third order Chebyshev filter, G_2 is a third order inverse Chebyshev filter and the static nonlinearity is built using a diode

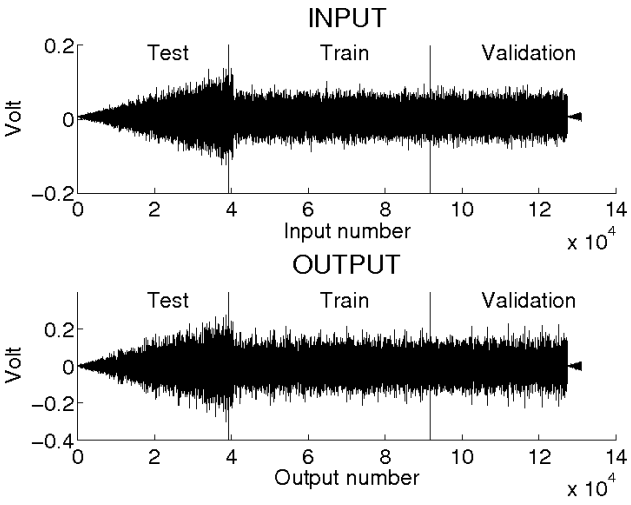


Fig. 1. Silverbox benchmark data set

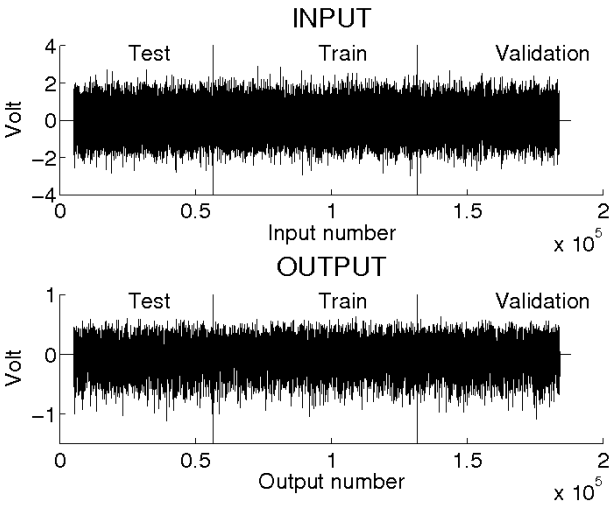


Fig. 3. Wiener-Hammerstein benchmark data set

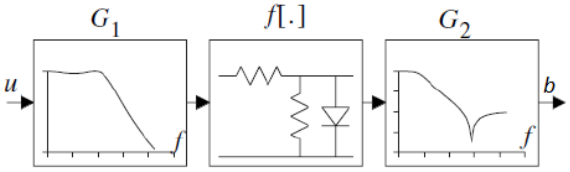


Fig. 2. Taken from [17]. Wiener-Hammerstein system consisting of a linear dynamic block G_1, a static non-linear block $f[\cdot]$ and a linear dynamic block G_2

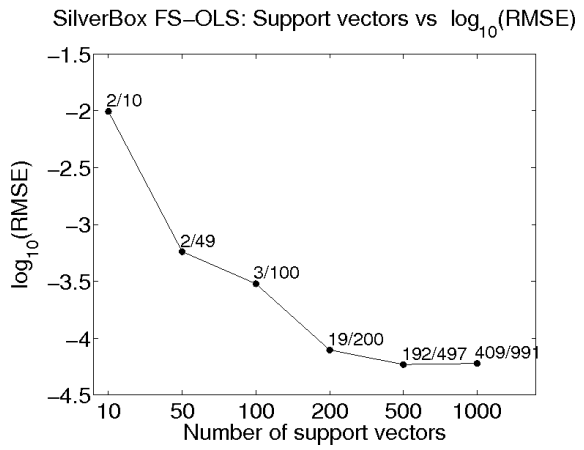
circuit. The measured input and output of the circuit are as shown on Figure 3. The data set consists of 188000 data points and was split evenly between test, validation and training sets. It can be found on *http://tc.ifac-control.org/1/1/Data%20Repository/sysid-2009-wiener-hammerstein-benchmark*

C. Truncation and generalization performance

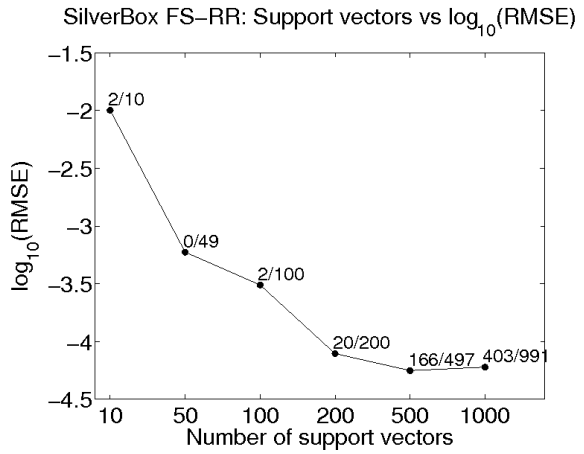
When the systems described in Section III are subjected to truncation, the general result obtained on the data sets used in this work is that the generalization performance decreases. This implies that if only the generalization performance is considered, the models should either remain unchanged or the truncation should be very minor in order to avoid the decrease in generalization performance. However, if a compromise between generalization performance and complexity is allowed, the situation changes dramatically. This can be seen in Figures 4 and 5 where a 10% in decreased generalization performance is allowed (i.e. the best generalization performance value is multiplied by 1.1 and this value is used as a tolerance threshold). In Figures 6 and 7 the resulting selection (i.e. with the 10% threshold) is represented by the diamond shaped markers. As can be seen, the more support values the system uses, the greater the reduction of singular values that can be achieved. Note that this holds for both data sets and for both FS-OLS and FS-RR methods. This behavior already suggests that the effective degrees of freedom can be greatly reduced if a small compromise of the generalization performance is allowed. This idea will be developed in section IV-D.

D. Effective number of parameters

The definitions in section III-C allow the representation of the effective number of degrees of freedom (given the different possible truncations) versus the generalization performance of the model. Figures 8 to 9 illustrate these results. Note that in this case, not only a good generalization performance is desired, but also a model with a reduced complexity. A compromise between both of them must be achieved. The lines suggest a possibly good choice for this compromise. To draw them, the axes are rescaled to be the same scale and the point with the minimum combined distance to the vertical axis and the lowest error in the rescaled axes is chosen. The rescaling is done to give the same relevance to both axes. The line is then drawn with the axes in their original scale and the graphs show that in these cases, it is possible indeed to greatly reduce the effective number of degrees of freedom without much loss of generalization performance.

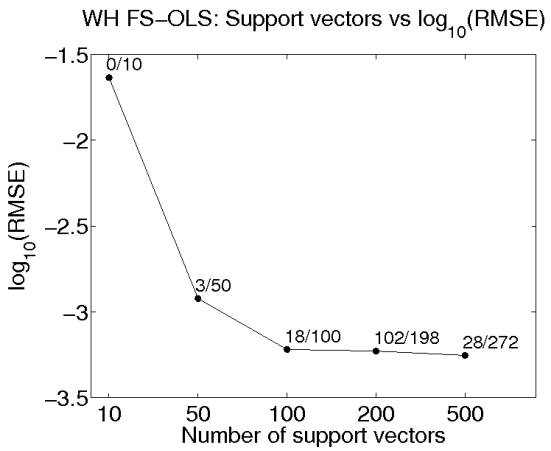


(a) FS-OLS

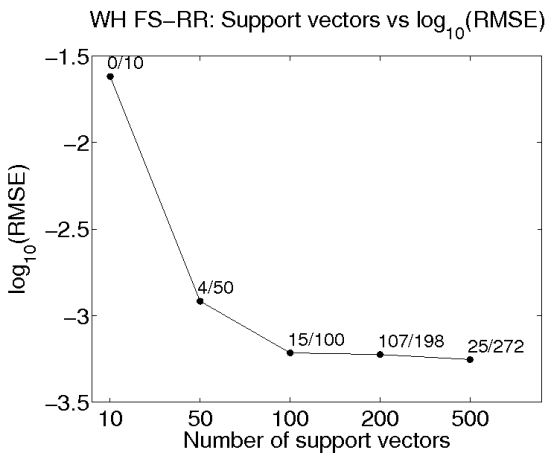


(b) FS-RR

Fig. 4. Test set performance vs. Number of support vectors on the Silverbox benchmark data set. At each point the relation between the number of singular values truncated and the total number of singular values is displayed.



(a) FS-OLS



(b) FS-RR

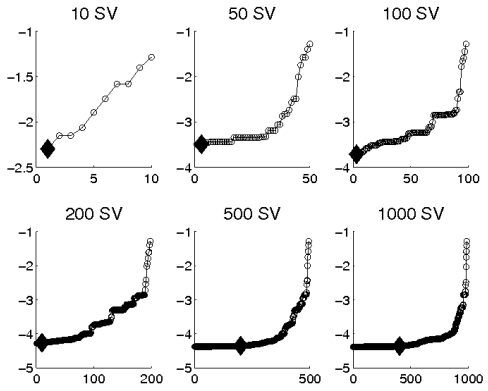
Fig. 5. Test set performance vs. Number of support vectors on the Wiener-Hammerstein benchmark data set. At each point the relation between the number of singular values truncated and the total number of singular values is displayed.

V. Discussion

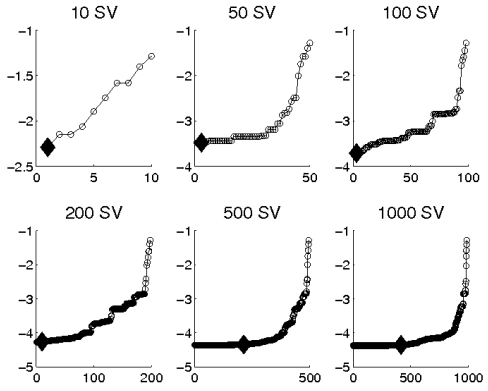
It has been shown in section IV-C that when applying SVD truncation schemes for Fixed-Size kernel models, in principle a significant reduction of support vectors is not to be expected if the generalization performance is to be maximized. However, if a trade-off between generalization performance and complexity is allowed, a significant truncation of the singular values of the $\boldsymbol{Q}$ matrix can be made. Furthermore, it has been shown that the complexity of the system, in terms of the effective degrees of freedom, can be greatly reduced through singular value truncation without a big impact on the generalization performance.

The results presented are relevant as they demonstrate that when employing Fixed-Size kernel models, it is possible to obtain models with highly reduced complexity when SVD truncation schemes are applied. However, those models will have a small reduction on generalization performance. This is desirable when the identified model is used e.g. for control purposes and when parsimonious models are preferred [12] and [9].

These findings are in line with [12], [14] and [18] as they illustrate that indeed the effective degrees of freedom for a Fixed-Size kernel model can greatly differ from the number of parameters of the system. In other words, the effective degrees of freedom can be much smaller than the number of support vectors in the Fixed-Size models.

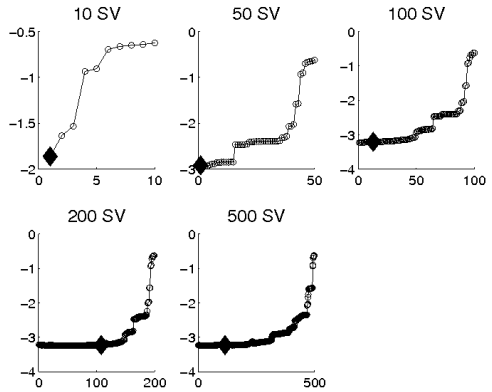


(a) FS-OLS

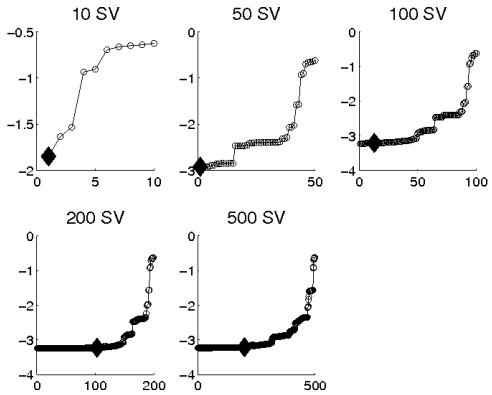


(b) FS-RR

Fig. 6. Compromise of up to 10% of test set performance for reduced complexity in the Silverbox benchmark data set. Horizontal axis represents the number of Singular Values eliminated. Vertical axis represents the test performance ($\log_{10}(RMSE)$).



(a) FS-OLS



(b) FS-RR

Fig. 7. Compromise of up to 10% of test set performance for reduced complexity in the Wiener-Hammerstein benchmark data set. Horizontal axis represents the number of Singular Values eliminated. Vertical axis represents the test performance ($\log_{10}(RMSE)$).

VI. Conclusions

In this paper we have considered different truncation schemes for fixed-size Kernel models based on SVD. It has been shown that if a compromise between generalization performance and complexity is allowed, the effective degrees of freedom of the underlying system can be greatly reduced on Fixed-Size kernel models without much loss of generalization performance.

FS-OLS and FS-RR methods have shown to very efficiently reduce the effective degrees of freedom of Fixed-Size kernel models under a SVD truncation scheme.

The methods presented have been successfully applied on two well-known benchmark data sets in system identification: the Wiener-Hammerstein and Silverbox data sets where similar and consistent results were obtained.

Possible future work may explore related methods for other possible model structures.

References

[1] Bishop C.M., *Neural Networks for Pattern Recognition*. Oxford University Press, Inc. New York, NY, USA, 1995.

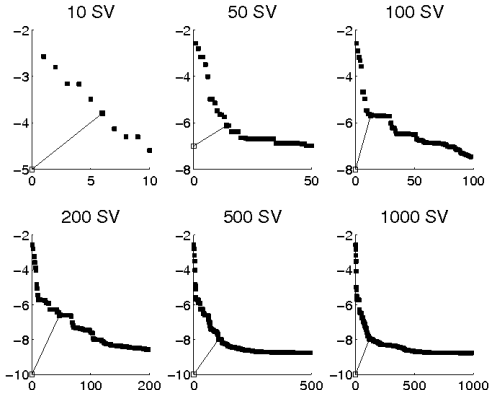
[2] De Brabanter K., De Brabanter J., Suykens J.A.K., De Moor B., "Approximate Confidence and Prediction Intervals for Least Squares Support Vector Regression" *IEEE Transactions on Neural Networks*, vol. 22, no. 1, Jan. 2011, pp. 110-120.

[3] De Brabanter K., Dreesen P., Karsmakers P., Pelckmans K., De Brabanter J., Suykens J.A.K., De Moor B., "Fixed-Size LS-SVM Applied to the Wiener-Hammerstein Benchmark", *Proc. of the 15th IFAC Symposium on System Identification (SYSID 2009)*, Saint-Malo, France, Jul. 2009, pp. 826-831.

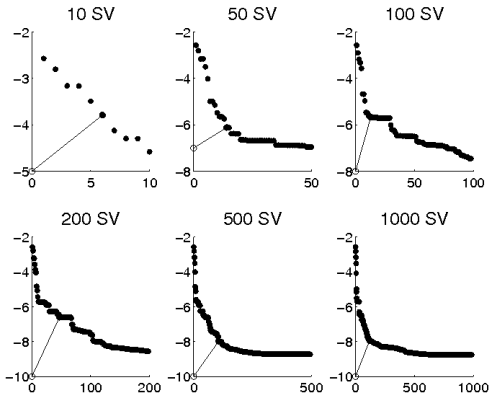
[4] Espinoza M., Pelckmans K., Hoegaerts L., Suykens J.A.K., De Moor B., "A comparative study of LS-SVMs applied to the Silver box identification problem", *Proc. of the 6th IFAC Symposium on Nonlinear Control Systems (NOLCOS 2004)*, Stuttgart, Germany, Sep. 2004.

[5] Espinoza M., Suykens J.A.K., De Moor B., "Kernel Based Partially Linear Models and Nonlinear Identification", *IEEE Transactions on Automatic Control*, Special Issue on System Identification, vol. 50, no. 10, Oct. 2005, pp. 1602-1606.

[6] Espinoza M., Suykens J.A.K., De Moor B., "Load Forecasting using Fixed-Size Least Squares Support Vector Machines", in *Computational Intelligence and Bioinspired Systems*, (Cabestany J., Prieto A., and Sandoval F., eds.), Proceedings

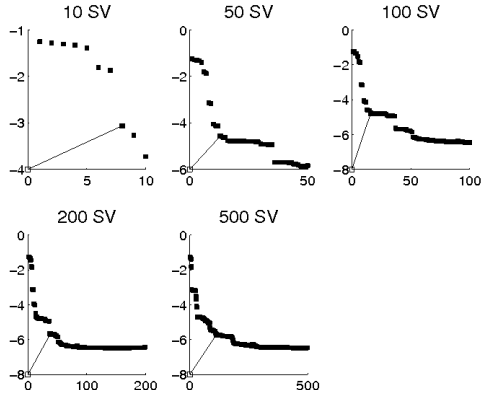


(a) FS-OLS

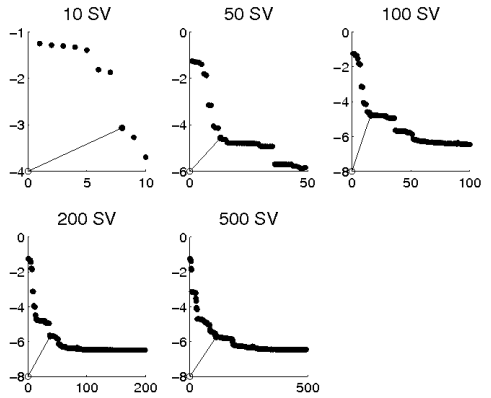


(b) FS-RR

Fig. 8. Test set performance vs EDF on the Silverbox benchmark data set for different fixed sizes. Horizontal axes represent the number of remaining effective degrees of freedom after truncation (i.e. $tr(H)$). The vertical axes represent the test set performance ($\log_{10}(RMSE)$).



(a) FS-OLS



(b) FS-RR

Fig. 9. Test set performance vs EDF for RR on the Wiener-Hammerstein benchmark data set for different fixed sizes. Horizontal axes represent the number of remaining effective degrees of freedom after truncation (i.e. $tr(H)$). The vertical axes represent the test set performance ($\log_{10}(RMSE)$).

of the 8th International Work-Conference on Artificial Neural Networks, vol. 3512 of Lecture Notes in Computer Science, Springer-Verlag, 2005, pp. 1018-1026.

[7] Falck T., Dreesen P., De Brabanter K., Pelckmans K., De Moor B., Suykens J.A.K., "Least-Squares Support Vector Machines for the Identification of Wiener-Hammerstein Systems", *Control Engineering Practice*, vol. 20, no. 11, Nov. 2012, pp. 1165-1174.

[8] Girolami M., "Orthogonal Series Density Estimation and the Kernel Eigenvalue Problem", in *Neural Computation* 14(3), 669-688, 2003

[9] Ljung L., *System Identification: Theory for the user (2nd Ed.)*. Prentice Hall, New Jersey, 1999.

[10] MacKay D.J.C., "Bayesian Interpolation", *Neural Computation* 4 (3): 415447. 1992.

[11] Mallows C.L., "Some comments on C_p", *Technometrics*, 15:661-6675, 1973.

[12] Marconato, A., Schoukens M., Rolain Y., Schoukens J., "Study of the Effective Number of Parameters in Nonlinear Identification Benchmarks", *52nd IEEE Conference on Decision and Control*, Florence, Italy, December 10-13, 2013, pp.4308-4313.

[13] Mercer J., "Functions of positive and negative type and their connection with the theory of integral equations", *Philos. Trans. Roy. Soc.*. 209, 415-446. London, 1909.

[14] Moody J.E., "The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems", In *NIPS*, Denver, Colorado, USA, 1991.

[15] Nyström E.J., "Über die praktische Auflösung von Integralgleichungen mit Anwendungen auf Randwertaufgaben", *Acta Mathematica*, 54:185-204, 1930.

[16] Schoukens J., Nemeth G., Crama P., Rolain Y., Pintelon R., "Fast Approximate Identification of Nonlinear Systems", *Automatica*, 39(7), 2003.

[17] Schoukens J., Suykens J., Ljung L., "Wiener-Hammerstein Benchmark", In *15h IFAC Symposium of System Identification*, Saint Malo, France, 2009.

[18] Spiegelhalter D.J., Best N.G., Carlin B.P., "Bayesian measures of model complexity and fit", *Journal of the Royal Statistical Society, Series B*, 2002.

[19] Suykens J.A.K., Van Gestel T., De Brabanter J., De Moor B., Vandewalle J., *Least Squares Support Vector Machines*, World Scientific Publishing Co., Pte, Ltd. (Singapore), (ISBN : 981-238-151-1), 2002.

[20] Vapnik V., *Statistical Learning theory*. Wiley, New-York, 1998.

[21] Williams C.K.I., Seeger M., "Using the Nyström method to speed up kernel machines", In T.K. Leen, T.G. Dietterich, and V. Tresp (Eds.) *Advances in Neural Information Processing Systems*, 13, 682-688, MIT Press. 2001.